Sentence Alignment Method Based on Maximum Entropy Model Using Anchor Sentences

Chao Che[1], Wenwen Guo[1], and Jianxin Zhang[1]

[1] Key Laboratory of Advanced Design and Intelligent Computing (Dalian University), Ministry of Education, Dalian
chechao101@163.com

Abstract. The paper proposes a sentence alignment method based on maximum entropy model using anchor sentences to align ancient and modern Chinese sentences in historical classics. The method selects the sentence pairs with the same phrases at the beginning or the end of the sentence or with the same time phrases as anchor sentence pairs, which are employed to divide the paragraph into several sections. Then, the sentences in each section are aligned using dynamic programming algorithm according to the entropy calculated by maximum entropy model. The maximum entropy model employs improved Chinese co-occurrence character feature, length feature and sentence alignment mode feature. The Chinese co-occurrence characters feature is improved by giving different weights to characters in different position based on the contribution to align sentences. In the experiment performed on ***ShiJi***, the precision and recall of the proposed method reaches 95.9% and 95.6% respectively, which outperforms other sentence alignment methods significantly.

Keywords: anchor sentences; maximum entropy model; Chinese co-occurrence character; sentence alignment

1 Introduction

History classics are the wisdom crystallization of the Chinese nation and the root of the modern culture. Translating history classics into English is one of direct and effective means to introduce the outstanding Chinese culture to the world. Due to the lack of ancient corpus and language process method of ancient Chinese, directly translating into English is very difficult. Therefore, we try to extract the term translation in the historical books using the modern Chinese as bridge, which may achieve good results with the help of rich resources of modern Chinese. The extraction method of term translation is based on parallel corpus between modern Chinese and ancient Chinese. Therefore, this paper presents a sentence alignment method, to provide the parallel corpus between ancient and modern Chinese for term translation of historical classics, taking ***Shiji*** for example.

The sentence alignment methods can be classified into three types: statistical method, lexical method and the combined method. The statistical method originally proposed by Brown[2] and Gale[3] is based on the fact: longer sentences in source language tend to be translated into longer sentences in target language, and shorter sentences tend to be translated into shorter sentences. The statistical method align sentence on the basis of sentence length. The lexical method use special symbols and special words (such as punctuations, mathematical symbols, the named entity etc), cognate information or lexicon to align the sentences [4,5,6]. The combined method employs both length and lexical information. For example, Wu [7] combined the sentence length with lexical information and achieved 92.1% accuracy on Chinese-English Hong Kong Hansard corpus.

Although the research of sentence alignment has been conducted for a long period of time, the research of sentence alignment between ancient and modern Chinese is still in initial stage. At present, the log-linear model and statistical method are employed to align sentences in ancient and modern corpus in [8,9]. In this paper, we employ maximum entropy model based anchor sentences to align ancient and modern sentences in historical classics. According to the characteristics of history books, we select anchor sentences and improve Chinese co-occurrence characters characteristic function.

2 A sentence alignment method based on maximum entropy model using anchor sentences

In the sentence alignment method, we firstly select two kinds of sentences as anchor [10,11], one type of sentence with the same phrases in begin or the end, the other type is *1-1* mode sentence pair containing the same time phrases. The anchor sentences are employed to divide the paragraph into several sections. Then, the entropy of ancient and modern sentences is calculated using maximum entropy model. At last, the sentences in each section are aligned by dynamic programming algorithm [12].

2.1 Selection of anchor sentences

The anchor sentences are implemented to control the alignment errors in a small range to prevent the spread of error. In theory, any kind of sentence pair can become anchors. But for the consideration of simplicity in practice, we only consider *1-1* mode sentences as anchor sentences [13], which accounts for a large majority of all the alignment

modes. Thus, we select two kinds of *1-1* mode sentence pairs in ancient and modern Chinese as the anchor sentences.

(1) Sentence pairs begin or end with the same phrases

Through the observation of ***Shiji*** in ancient and modern Chinese, we find that some aligned sentences have same phrases at the beginning and the end of the sentence, while the non-aligned sentences do not have the characteristic. Obviously the sentences with the same phrases at the beginning and the end can be used as anchor sentences. Two points should be noted: ① Ancient modal words does not make sense, so the modal words at the beginning and the end of ancient sentences are re-moved, including "维"(wei), "而"(er), "何"(he), "乎"(hu), etc. ②Some of the ancient Chinese words are replaced by the modern Chinese characters. Some characters have the same meaning in both modern Chinese and ancient Chinese, the characters need to be replaced. For example, "於是"(yushi) in ancient Chinese is replaced by"于是"(yushi) in modern, "曰"(yue) is replaced by "说"(shuo).

(2) Sentence pairs with the same time phrases

Shiji is a biographical history book, which records the history by biography. Each biography is recorded in chronological order, so there are a large number of time phrases in the sentence of ***Shiji***. Time phrases containing "年"(nian) is more typical while the time phrases in ancient and modern Chinese are the same, sentence pairs with the same time phrases can be regarded as anchor sentences. In our training corpus, there are 341 pairs of sentences containing the same time phrases among 2798 pairs, some of which are shown in Table 1, (β is the ratio of Chinese co-occurrence characters and the number of ancient sentence character)

Table 1. Sentences with the same time phrases

Ancient Chinese	Modern Chinese	β
出子六年三父等复共令人贼杀出子	出子六年三父等人又合伙派人谋杀出子	0.73
四年晋伐秦取少梁	四年晋伐秦占领少梁	0.875
景公四年晋栾书弒其君厉公	景公四年晋国的栾书杀其国君厉公	0.92
五年晋卿中行范氏反晋	五年晋卿中行氏范氏反叛晋国	1
十三年初有史以纪事民多化者	十三年开始设史官记事人民越来越开化	0.58

The sentence pairs aligned to each other usually have big β. There are little anchor sentences with the small β. In the training corpus, there are only 10 β of sentences less than or equal to 0.6. So the pair of sentences with $\beta \leq 0.6$ can't been regarded as anchor. We find the anchor sentences with the same time phrases in following steps:

① Calculate the sentence number rate $\alpha = M/N$, in which, N is the number of sentences in ancient paragraph, M is the number of sentences in corresponding modern paragraph.

② Assume the *i-th* ancient sentences contains time phrases with " 年"(nian), *j-th* modern sentence is the possible translation corresponds to the *i-th* ancient sentences. $j = \alpha * i + k$, sets $k = [\pm 1, 0, \pm 2]$ from experience.

③ Calculate the value of β for each sentence pairs and regard the maximal value as $\beta_{\max}$. Remove the sentence pairs with $\beta_{\max} \leq 0.6$.

④ Determine whether the sentence pair is *1-1* alignment mode. Seeking all possible entropy with different alignment modes. If the entropy of $s_i - t_u$ with *1-1* mode is the optimal solution, then $s_i - t_u$ can be regarded as anchor sentences. The sentence pairs of different alignment modes needed to be calculated are shown in Table 2.

Table 2. Sentence combinations needed seek entropy

Alignment mode	Sentence combinations	
1–1	$s_i - t_u$	
1–2	$s_i - t_{u-1}, t_u$	$s_i - t_u, t_{u+1}$
2–1	$s_{i-1}, s_i - t_u$	$s_i, s_{i+1} - t_u$
2–2	$s_{i-1}, s_i - t_{u-1}, t_u$	$s_{i-1}, s_i - t_u, t_{u+1}$
	$s_i, s_{i+1} - t_{u-1}, t_u$	$s_i, s_{i+1} - t_u, t_{u+1}$

⑤ Repeat above step， find anchor sentences until the end of the paragraph.

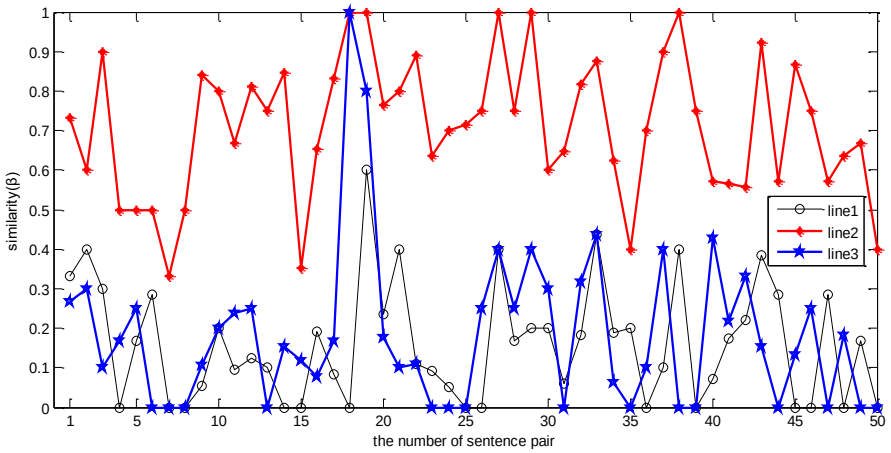
2.2 Maximum Entropy Model

For bilingual sentence alignment, the integration of multiple characteristic functions can get good results. The maximum entropy model makes use of sentence length, alignment mode, Chinese co-occurrence character. Sentence alignment process is to compute the entropy of ancient and modern sentences. And the formula is shown in Equation (1).

$$\begin{aligned}\arg\max e(s,t) &= \arg\max[\lambda_m h_m(s,t)] \\ &= \arg\max[\lambda_1 L(s,t) + \lambda_2 M(s,t) + \lambda_3 H(s,t)]\end{aligned} \quad (1)$$

In which, $L(s,t)$ is the length feature function, $M(s,t)$ is the alignment mode feature function, $H(s,t)$ is the Chinese co-occurrence characters feature function.

Chinese co-occurrence characters. Through observation on corpus, we find that two sentences with more co-occurrence Chinese characters, the greater the chance of alignment. In the real corpus, there are some the same Chinese characters in ancient sentence and the corresponding modern sentence, there are also the co-occurrence of Chinese characters in the surrounding sentences. If the Chinese co-occurrence characters are used in the sentence alignment, it is in ancient sentence and the corresponding modern sentence far more than in the surrounding sentence pairs. In order to verify whether the Chinese co-occurrence characters can be used, we have selected 50 ancient sentences from manual alignment corpus to calculate similarity β, the results shown in Fig.1.



line1: The similarity with the former one. line2: The similarity with the aligned one

line3: The similarity with the after one

Fig. 1. The similarity of Chinese co-occurrence characters in align and unaligned sentences

In Fig. 1, line 1 denotes the similarity between ancient sentence and the sentence before its corresponding modern Chinese translation, line 2 is the similarity between ancient sentence and its corresponding translation, line 3 denotes the similarity between ancient and the sentence after its corresponding translation. From Fig. 1 we can see the fact that the similarities between the aligned sentences is much bigger than the unaligned ones. Wang [14] puts forward an algorithm to calculate the co-occurrence character feature in the alignment between China and Japan, shown as formula (2). We improve the formula by giving different weights to characters in different position

based on the contribution to sentences alignment especially begin and the end of a sentence.

$$H(s,t)=\begin{cases} 0 & 0-1 or 1-0 \\ h_1(s_i,t_u:0,0) & 1-1 \\ h_1(s_i,t_u:0,0)+h_2(s_i,t_u;s_j,0) & 2-1 \\ h_1(s_i,t_u:0,0)+h_3(s_i,t_u;0,t_v) & 1-2 \\ h_1(s_i,t_u:0,0)+h_1(0,0;s_j,t_v) & 2-2 \end{cases} \tag{2}$$

Formula (3) indicates the similarity between s_i and t_u. Formula (4) is the similarity increased by the characters that appear in s_j and do not appear in s_i. Formula (5) indicates the similarity increased by the characters that appear in t_v and do not appear in t_u.

$$h_1\left(s_i,t_u:0,0\right)=\frac{c\left(s_i\bigcap t_u\right)}{\min\left(c\left(s_i\right),c\left(t_u\right)\right)} \tag{3}$$

$$h_2\left(s_i,t_u:s_j,0\right)=\frac{c\left(s_j\bigcap t_u\right)-c\left(s_i\bigcap s_j\bigcap t_u\right)}{c\left(s_j\right)} \tag{4}$$

$$h_3(s_i,t_u:0,t_v)=\frac{c\left(s_i\bigcap t_v\right)-c(s_i\bigcap t_u\bigcap t_v)}{c(t_v)} \tag{5}$$

$c(s_i)$ denotes the character number in ancient sentence s_i, $c(t_u)$ denotes the number in modern sentence t_u, $c\left(s_i\bigcap t_u\right)$ denotes the number of the co-occurrence characters in both ancient sentence s_i and modern sentence t_u.

There are many nouns co-occur in both ancient and modern corpus, and most co-occurrence words appear at the beginning or at the end of the ancient sentence and its modern translation sentence. According to the different contribution to sentence alignment, words at different positions are gave different weights. Formula (3) are improved as formula (6).

$$\begin{aligned} h_1\left(s_i,t_u:0,0\right)=&\frac{\lambda_1\left[c\left(s_i\bigcap t_u\right)-c_b(s_i\bigcap t_u)-c_e(s_i\bigcap t_u)\right]}{\min(c(s_i),c(t_u))} \\ &+\frac{\lambda_2 c_b\left(s_i\bigcap t_u\right)+\lambda_3 c_e\left(s_i\bigcap t_u\right)}{\min(c(s_i),c(t_u))} \end{aligned} \tag{6}$$

$c_b(s_i\bigcap t_u)$ indicates the number of co-occurrence characters at the beginning of the sentence. $c_e(s_i\bigcap t_u)$ denotes the number of co-occurrence characters at the end of sentence.

Sentence length. The alignment method employs sentence lengths based on the fact: longer sentences in source language tend to be translated into longer sentences in target language, and shorter sentences tend to be translated into shorter sentences [2]. In the study of the Gale and Church [3], each character in a language will corresponds to certain random number of characters in another language. The character number is independent identically distributed, complying with normal distribution.

The sentence length is denoted by the number of Chinese characters. In our experiment, l_2 is the length of the modern sentence, l_1 indicates the length of the ancient sentence. We fit the length ratio l_2/l_1 in the corpora of ancient and modern Chinese sentences to normal distribution, fitting results is shown in Fig.2.

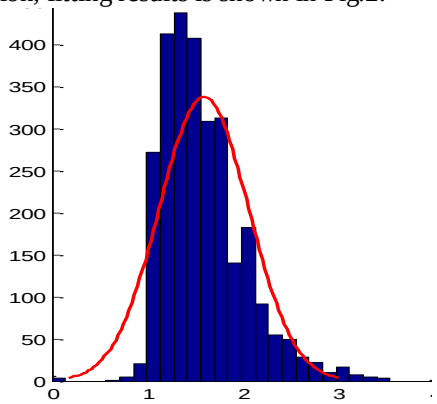


Fig.2. Normal distribution curve of length ratio

In Fig.2, we find the length ratio l_2/l_1 obeys normal distribution $c \approx 1.588$ $s^2 = 2.007$. Given $\delta = (l_2 - l_1 \cdot c)/\sqrt{l_1 \cdot s^2}$, δ meets the standard normal distribution[15].

Sentence length feature is computed using the formula (7) [8].

$$L(s,t) = -100 \times \log_2\left(1 - 1\Big/\sqrt{2\pi}\int_{-\infty}^{\delta} e^{-\frac{z^2}{2}} dz\right) \quad (7)$$

Table 3. Alignment Mode Statistics

mode(m-n)	frequency	probability
1-0 or 0-1	4	0.14%
1-1	2522	90.14%
1-2	132	4.72%
2-1	112	4.00%
2-2	8	0.29%
others	20	0.71%

Alignment Mode. Analyzing the parallel corpora of ***Shiji***, we get the probability of different alignment modes in Table 3.

In the alignment, we consider only the six modes including *1-0, 0-1, 1-1, 1-2, 2-1, 2-2*. If m ancient Chinese sentences align n modern Chinese sentences, we call the alignment as m-n alignment mode. The alignment mode feature $M(s,t)$ is calculated using the formula (8) [9].

$$M(s,t) = -100\log(\Pr(m-n)) \tag{8}$$

3 Experiments and Analysis

3.1 Experiment settings

The parallel corpus used in this experiment is composed of the ancient Chinese and modern Chinese translation of five basic annals from ***ShiJi,*** including ***Basic Annals of Qin***, ***the Basic Annals of the First Emperor of the Qin***, ***the Basic Annals of Hsiang Yu*** , ***the Basic Annals of Emperor Kao-tsu*** and ***the Basic Annals of Empress Lü*** as corpus[16]. 2798 sentences pairs are selected as training corpus and 1341 sentences pairs are selected as test corpus.

The experimental results are evaluated by precision (P), recall ratio(R), and F measure. R, P and F is defined as (9), (10) and (11) respectively.

$$P = A/(A+B) \tag{9}$$

$$R = A/(A+C) \tag{10}$$

$$F = 2\times(RP/(R+P)) \tag{11}$$

In which, A is the number of sentence correct aligned, B is the number of sentence wrong aligned, C is the number of sentence not correct aligned.

In the experiment, we employ different combination of features to do sentence alignment and the results are showed in Table 4. CC denotes Chinese co-occurrence characters feature presented by Wang [14], improved Chinese co-occurrence characters feature proposed by this paper is denoted by ICC. A denotes anchor sentences feature.

3.2 Analysis

Many factors affect the alignment results. According to the characteristics of the method, we analyze the results from the following three aspects.

Table 4. The experimental results using different features

length	mode	CC	ICC	A	P	R	F
+					78.6%	51.2%	62.0%
		+			80.3%	52.8%	63.7%
			+		84.2%	56.4%	67.5%
+	+				89.8%	89.0%	89.4%
	+		+		90.2%	90.0%	90.1%
+	+		+		93.5%	93.3%	93.4%
+	+		+	+	95.9%	95.6%	95.7%

(1) Anchor sentences

The anchor sentences, which are the sentence pairs with the same phrase at the beginning and the end of the sentences, have similar effects with the improved Chinese co-occurrence characters feature. The anchor sentences require aligned entirely correct, so the accuracy of using anchor sentence is higher than using Chinese co-occurrence characters.

Sentences with time phrases describe events including person names, place names or time etc. These phrases often have the same representation in ancient and modern Chinese, and there are more co-occurrence Chinese characters. Therefore similarity can be identified as a measure to determine anchor sentences.

One disadvantage of the anchor sentences is the uneven distribution. Therefore, the anchor sentences take effect obviously in some paragraphs, while do not work in other paragraphs. For example, the paragraph from “十三年，向寿伐韩，取武始” to “五十年十月，武安君白起有罪，为士伍，迁阴密”in ***Basic Annals of Qin***, has 35 time parses with “年”(nian), including 77 sentences. Clearly the role of anchor sentence in this paragraph will be very obvious.

(2) Chinese co-occurrence character

As shown in Table 4, when only a feature is used, the co-occurrence characters features obtains the best result. The improved co-occurrence characters feature that gives different weights to co-occurrence characters in different position perform even better. When only using co-occurrence characters feature, the alignment is conducted according to the similarity between the two sentences. Chinese characters appear relatively random, so the similarity in different sentences pairs may be the same or not very different. However, if the position of co-occurrence character can be considered, accuracy is increased by 3.9% as shown in Table 4. The improved accuracy mainly from the

sentences with same characters at the beginning or the end of the sentence as shown in Table 5.

Table 5. Aligned sentences with same characters at the beginning or the end of sentences

Ancient	Modern sentences
女脩织，. . . 生子大业。	女脩织布时，. . . 生下儿子大业。
其玄孙曰费昌，. . . 或在夷狄	费氏的玄孙叫费昌。. . . ，有些住在夷狄。
项梁前使项 羽别攻襄城，	项梁在这之前派项羽另率一军攻打襄城

(3) Length Feature

The length feature intends to match sentences whose length ratio is around 1.588, and it is easy to cause a mismatch for distribution on both ends of the peak. Due to the symmetry of normal distribution, the length feature cannot work well in both wings of the normal distribution, slightly different will cause the mismatch.

4 Conclusion

In this paper, we conduct a preliminary study of sentence alignment in historical classics, and achieve satisfactory results using the maximum entropy model based on the anchor sentences. Since the sentence alignment method is based on the characteristic of Chinese characters, the application of the method has many limitations. In the future research, we will study a widely used alignment model not limited to ancient Chinese.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 61402068) and Support Program of Outstanding Young Scholar in Liaoning Universities. (No. LJQ2015004).

References:

1. Sima,Q. (Han dynasty): Shiji. Zhong Hua Book Company, Beijing (2006)
2. Brown , P. F. , Lai , J. C. , Mercer , R. L.: Aligning sentences in parallel corpora . In : Proceedings of 29th Annual Conference of the Association for Computational Linguistics，ACL 1991 , Stroudsburg, PA, USA, pp. 169-176 (1991)
3. Gale , W.A. , Church, K.W. : A Program for Aligning Sentences in Bilingual Corpora. In: Proceedings of 29th Annual Conference of the Association for Computational Linguistics，MIT , MA, USA ,pp. 19(1) :75-102 (1993)

4. Kay, M.，Roscheisen, M. : Text-translation Alignment. Computational Linguistics, MIT, MA, USA, 19(1):121-142 (1993)
5. Chen, S. F.: Aligning sentences in bilingual corpora using lexical information. In: Proceedings of 31st Annual Meeting of the Association for Computational Linguistics，pp. 9-16, ACL, Stroudsburg, (1993)
6. Simard, M., Foster,G.F., Isabelle, P.: Using cognates to align sentences in bilingual corpora. In :Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative research, pp. 1071-1082, IBM Press, Indianapolis (1993)
7. Wu, D. K.: Aligning a parallel English-Chinese corpus statistically with lexical criteria. In: Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics, pp. 80-87, ACL, Stroudsburg, USA , (1994)
8. Liu, Y., Wang, N.: Research on Classical and Modern Chinese Sentence Alignment. Computer Applications and Software, 30(11):127-130 (2013)
9. Lin, Z.: Alignment for Ancient-Modern Chinese Bi-text. Beijing University of Posts and Telecommunications, Beijing (2007)
10. Zhou,Y.Q.: Maximum Entropy Method and its Applications in Natural Language Processing. Fudan University , Shanghai(2004)
11. Berger, A. L., Pietra, V. J. D., Pietra, S. A. D.: A Maximum Entropy Approach to Natural Language Processing. In: Proceedings of the 34th Annual Conference of the Association for Computational Linguistics, pp. 39-71, ACL, Stroudsburg (1996)
12. Cormen, T. H., .Leiserson, C. E., Rivest, R. L., Stein, C.: Introduction to Algorithms. pp. 323-369, MIT Press, Cambridge (2001)
13. Tian, S.W., TURGUN. I, YU. L, et al.: Chinese-Uyhur Sentence Alignment Based on Hybrid Strategy. Computer Science, 37(4) : 215-218 (2010)
14. Wang, X.J., Ren, F.J.: Chinese-Japanese Clause Alignment. In: 6th International Conference of Computational Linguistics and Intelligent Text Processing , pp.400-412, Springer, Heidelberg (2005)
15. Zhu, G.J., Guo, D.W, Liu, X.: Probability Theory and Mathematical Statistics. National Defence Industry Press, Beijing (2010)
16. Watson B.: Records of the Grand Historian: Qin Dynasty. Chinese University of Hong Kong Press, Hong Kong (1993)